NATIONAL QUALITY FORUM—Measure Testing (subcriteria 2a2, 2b2-2b7)

Measure Number (*if previously endorsed*)**:** 0061
Measure Title: Comprehensive Diabetes Care: Blood Pressure Control <140/90 mm Hg
Date of Submission: 12/5/2014
Type of Measure:

☐ Composite – ***STOP – use composite testing form***	☒ Outcome (*including PRO-PM*) Intermediate Outcome
☐ Cost/resource	☐ Process
☐ Efficiency	☐ Structure

Instructions

- Measures must be tested for all the data sources and levels of analyses that are specified. ***If there is more than one set of data specifications or more than one level of analysis, contact NQF staff*** about how to present all the testing information in one form.
- **For all measures, sections 1, 2a2, 2b2, 2b3, and 2b5 must be completed.**
- **For outcome and resource use measures**, section **2b4** also must be completed.
- If specified for **multiple data sources/sets of specificaitons** (e.g., claims and EHRs), section **2b6** also must be completed.
- Respond to all questions as instructed with answers immediately following the question. All information on testing to demonstrate meeting the subcriteria for reliability (2a2) and validity (2b2-2b6) must be in this form. An appendix for *supplemental* materials may be submitted, but there is no guarantee it will be reviewed.
- If you are unable to check a box, please highlight or shade the box for your response.
- Maximum of 20 pages (*incuding questions/instructions;* minimum font size 11 pt; do not change margins). ***Contact NQF staff if more pages are needed.***
- Contact NQF staff regarding questions. Check for resources at Submitting Standards webpage.

Note: The information provided in this form is intended to aid the Steering Committee and other stakeholders in understanding to what degree the testing results for this measure meet NQF's evaluation criteria for testing.

2a2. Reliability testing [10] demonstrates the measure data elements are repeatable, producing the same results a high proportion of the time when assessed in the same population in the same time period and/or that the measure score is precise. For **PRO-PMs and composite performance measures**, reliability should be demonstrated for the computed performance score.

2b2. Validity testing [11] demonstrates that the measure data elements are correct and/or the measure score correctly reflects the quality of care provided, adequately identifying differences in quality. For **PRO-PMs and composite performance measures**, validity should be demonstrated for the computed performance score.

2b3. Exclusions are supported by the clinical evidence; otherwise, they are supported by evidence of sufficient frequency of occurrence so that results are distorted without the exclusion; [12]
AND
If patient preference (e.g., informed decisionmaking) is a basis for exclusion, there must be evidence that the exclusion impacts performance on the measure; in such cases, the measure must be specified so that the information about patient preference and the effect on the measure is transparent (e.g., numerator category computed separately, denominator exclusion category computed separately). [13]

2b4. For outcome measures and other measures when indicated (e.g., resource use):

- **an evidence-based risk-adjustment strategy** (e.g., risk models, risk stratification) is specified; is based on patient factors that influence the measured outcome (but not factors related to disparities in care or the quality of care) and are present at start of care; [14,15] and has demonstrated adequate discrimination and calibration

OR

- rationale/data support no risk adjustment/ stratification.

2b5. Data analysis of computed measure scores demonstrates that methods for scoring and analysis of the specified measure allow for **identification of statistically significant and practically/clinically meaningful** [16] **differences in performance**;

OR

there is evidence of overall less-than-optimal performance.

2b6. If multiple data sources/methods are specified, there is demonstration they produce comparable results.

2b7. For **eMeasures, composites, and PRO-PMs** (or other measures susceptible to missing data), analyses identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias.

Notes

10. Reliability testing applies to both the data elements and computed measure score. Examples of reliability testing for data elements include, but are not limited to: inter-rater/abstractor or intra-rater/abstractor studies; internal consistency for multi-item scales; test-retest for survey items. Reliability testing of the measure score addresses precision of measurement (e.g., signal-to-noise).

11. Validity testing applies to both the data elements and computed measure score. Validity testing of data elements typically analyzes agreement with another authoritative source of the same information. Examples of validity testing of the measure score include, but are not limited to: testing hypotheses that the measures scores indicate quality of care, e.g., measure scores are different for groups known to have differences in quality assessed by another valid quality measure or method; correlation of measure scores with another valid indicator of quality for the specific topic; or relationship to conceptually related measures (e.g., scores on process measures to scores on outcome measures). Face validity of the measure score as a quality indicator may be adequate if accomplished through a systematic and transparent process, by identified experts, and explicitly addresses whether performance scores resulting from the measure as specified can be used to distinguish good from poor quality.

12. Examples of evidence that an exclusion distorts measure results include, but are not limited to: frequency of occurrence, variability of exclusions across providers, and sensitivity analyses with and without the exclusion.

13. Patient preference is not a clinical exception to eligibility and can be influenced by provider interventions.

14. Risk factors that influence outcomes should not be specified as exclusions.

15. Risk models should not obscure disparities in care for populations by including factors that are associated with differences/inequalities in care, such as race, socioeconomic status, or gender (e.g., poorer treatment outcomes of African American men with prostate cancer or inequalities in treatment for CVD risk factors between men and women). It is preferable to stratify measures by race and socioeconomic status rather than to adjust out the differences.

16. With large enough sample sizes, small differences that are statistically significant may or may not be practically or clinically meaningful. The substantive question may be, for example, whether a statistically significant difference of one percentage point in the percentage of patients who received smoking cessation counseling (e.g., 74 percent v. 75 percent) is clinically meaningful; or whether a statistically significant difference of $25 in cost for an episode of care (e.g., $5,000 v. $5,025) is practically meaningful. Measures with overall less-than-optimal performance may not demonstrate much variability across providers.

1. DATA/SAMPLE USED FOR ALL TESTING OF THIS MEASURE

Often the same data are used for all aspects of measure testing. In an effort to eliminate duplication, the first five questions apply to all measure testing. If there are differences by aspect of testing,(e.g., reliability vs. validity) be sure to indicate the specific differences in question 1.7.

1.1. What type of data was used for testing? (*Check all the sources of data identified in the measure specifications and data used for testing the measure. Testing must be provided for all the sources of data specified and intended for measure implementation.* ***If different data sources are used for the numerator and denominator, indicate N [numerator] or D [denominator] after the checkbox.***)

Measure Specified to Use Data From: (*must be consistent with data sources entered in S.23*)	**Measure Tested with Data From:**
☒ abstracted from paper record	☒ abstracted from paper record
☒ administrative claims	☒ administrative claims
☐ clinical database/registry	☐ clinical database/registry
☐ abstracted from electronic health record	☐ abstracted from electronic health record
☐ eMeasure (HQMF) implemented in EHRs	☐ eMeasure (HQMF) implemented in EHRs
☐ other: Click here to describe	☐ other: Click here to describe

1.2. If an existing dataset was used, identify the specific dataset (*the dataset used for testing must be consistent with the measure specifications for target population and healthcare entities being measured; e.g., Medicare Part A claims, Medicaid claims, other commercial insurance, nursing home MDS, home health OASIS, clinical registry*).

N/A

1.3. What are the dates of the data used in testing?

Health Plan Level (2014) HEDIS

Testing of performance measure score with beta binomial reliability was performed with data from measurement year 2013.

Testing of construct validity with Pearson's Correlation was performed using data from measurement year 2013.

Clinician/Practice Diabetes Recognition Program (2013)

Testing of performance measure score with beta binomial reliability was performed with data from measurement year 2012.

Testing of construct validity with Pearson's Correlation was performed using data from measurement year 2012.

1.4. What levels of analysis were tested? (*testing must be provided for all the levels specified and intended for measure implementation, e.g., individual clinician, hospital, health plan*)

Measure Specified to Measure Performance of: (*must be consistent with levels entered in item S.26*)	**Measure Tested at Level of:**
☒ individual clinician	☒ individual clinician
☒ group/practice	☒ group/practice
☐ hospital/facility/agency	☐ hospital/facility/agency
☒ health plan	☒ health plan
☐ other: Click here to describe	☐ other: Click here to describe

1.5. How many and which measured entities were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of measured entities included in the analysis (e.g., size, location, type); if a sample was used, describe how entities were selected for inclusion in the sample*)

Sample for Beta Binomial Testing and Construct Validity Testing

Measure score reliability and construct validity were calculated from U.S. HEDIS data that include 342 Commercial plans, 209 Medicaid plans and 489 Medicare plans. The sample included all plans submitting data to NCQA for HEDIS. The plans were geographically diverse and varied in size.

Measure score reliability and construct validity were calculated from the Diabetes Recognition Program that included 2,477 clinicians. This total includes 515 clinicians from 123 practices.

Systematic Assessment for Face Validity

This measure was tested for face validity with three expert panels. See additional information: Ad.1. Workgroup/Expert Panel in Measure Development for names and affiliation of expert panels:

1. Diabetes Measurement Advisory Panel includes 17 clinicians with expertise in endocrinology, podiatry, nephropathy, ophthalmology, cardiology, internal medicine/family medicine, pediatric endocrinology pharmacy and diabetes education. Other stakeholders include one consumer advocate and one health advisory consultant with expertise in diabetes treatment.
2. Cardiovascular Measurement Advisory Panel includes eight physicians and one nurse with expertise in cardiovascular health and quality measurement.
3. NCQA's Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 16 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

1.6. How many and which patients were included in the testing and analysis (by level of analysis and data source)? (*identify the number and descriptive characteristics of patients included in the analysis (e.g., age, sex, race, diagnosis); if a sample was used, describe how patients were selected for inclusion in the sample*)

Samples for Beta binomial and Construct Validity testing

Health Plan Level

In measurement year 2013, HEDIS measures covered more than 171 million people from 814 HMOs and 353 PPOs. Data are summarized at the health plan level and stratified by product line (i.e. commercial, Medicare, Medicaid). Below is a description of the sample. It includes the number of health plans included in HEDIS data collection and the average sample size used for reporting.

Product Line	Number of Plans	Average Denominator Size per Plan (using HEDIS Hybrid sampling methodology)
Commercial (HMO/PPO combined)	342	1,325
Medicaid	209	475
Medicare (HMO/PPO combined)	489	551

Diabetes Recognition Program

Clinicians submitting data to the Diabetes Recognition Program must submit data on a sample of their patient population with diabetes. The sample includes at least 25 different eligible patients with diabetes per clinician. This sampling method also applies to group practices (at least 25 different eligible patients with diabetes for each clinician per site). In 2013, a total of 120,887 patients were included in the sample.

1.7. If there are differences in the data or sample used for different aspects of testing (e.g., reliability, validity, exclusions, risk adjustment), identify how the data or sample are different for each aspect of testing reported below.

Reliability and construct validity of this measure were tested using the same sample. Validity was demonstrated through a systematic assessment of face validity. Per NQF instructions, the composition of each panel is described in question 1.5 above.

2a2. RELIABILITY TESTING

Note*: If accuracy/correctness (validity) of data elements was empirically tested, separate reliability testing of data elements is not required – in 2a2.1 check critical data elements; in 2a2.2 enter "see section 2b2 for validity testing of data elements"; and skip 2a2.3 and 2a2.4.*

2a2.1. What level of reliability testing was conducted? (*may be one or both levels*)

☐ **Critical data elements used in the measure** (*e.g., inter-abstractor reliability; data element reliability must address ALL critical data elements*)

☒ **Performance measure score** (e.g., *signal-to-noise analysis*)

2a2.2. For each level checked above, describe the method of reliability testing and what it tests (*describe the steps—do not just name a method; what type of error does it test; what statistical analysis was used*)

Performance Measure Score (Beta Binomial)

In order to assess measure precision in the context of the observed variability across accountable entities, we utilized the reliability estimate proposed by Adams (2009). The following is quoted from the tutorial which focused on provider-level assessment: "Reliability is a key metric of the suitability of a measure for [provider] profiling because it describes how well one can confidently distinguish the performance of one physician from another. Conceptually, it is the ratio of signal to noise. The signal in this case is the proportion of the variability in measured performance that can be explained by real differences in performance. There are three main drivers of reliability: sample size, differences between physicians, and measurement error. At the physician level, sample size can be increased by increasing

the number of patients in the physician's data as well as increasing the number of measures per patient." This approach is also relevant to health plans and other accountable entities.

Adams' approach uses a Beta-binomial model to estimate reliability; this model provides a better fit when estimating the reliability of simple pass/fail rate measures as is the case with most HEDIS® measures. The beta-binomial approach accounts for the non-normal distribution of performance within and across accountable entities. Reliability scores vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise or the individual accountable entity variance) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (across accountable entities).

Adams, J. L. The Reliability of Provider Profiling: A Tutorial. Santa Monica, California: RAND Corporation. TR-653-NCQA, 2009

2a2.3. For each level of testing checked above, what were the statistical results from reliability testing? (e.*g., percent agreement and kappa for the critical data elements; distribution of reliability statistics from a signal-to-noise analysis*)

Performance Measure Score (Beta Binomial):

Health Plan Level (Measurement Year 2013)

Below, we present health plan level data, which includes data submitted from 342 commercial plans, 209 Medicaid plans and 489 Medicare plans.

Commercial			Medicaid			Medicare		
Median	Overall	10th-90th	Median	Overall	10th-90th	Median	Overall	10th-90th
0.98	0.99	0.97-0.98	0.97	0.97	0.96-0.98	0.95	0.97	0.92-0.97

Diabetes Recognition Program (Measurement Year 2012)

Below, we present available statistics from the Diabetes Recognition Program (BP >=140/90 mm Hg Lower score=Better quality). We remind the panel that this is a voluntary program subject to self-selection and involves 2,477 individual clinicians. This total includes 515 clinicians from 123 practices.

Diabetes Recognition Program (Clinician and Practice)		
Median	Overall	10th-90th
0.63	0.70	0.41-0.90

2a2.4 What is your interpretation of the results in terms of demonstrating reliability? (i.*e., what do the results mean and what are the norms for the test conducted?*)

Performance Measure Score (Beta Binomial):

Reliability scores can vary from 0.0 to 1.0. A score of zero implies that all variation is attributed to measurement error (noise) whereas a reliability of 1.0 implies that all variation is caused by a real difference in performance (signal). Generally, a minimum reliability score of 0.7 is used to indicate sufficient signal strength to discriminate performance between accountable entities.

Health Plan Level

Testing suggests that this measure has very good reliability at the health plan level between 0.97 and 0.99. The 10-90th percentile distribution of health plan level reliability on this measure shows the vast majority of health plans met or exceeded the minimally accepted threshold of 0.7, and the majority of plans exceeded 0.8. Strong reliability is demonstrated since variances in these large populations are due to signal and not to noise.

Diabetes Recognition Program

Testing suggests that this measure has good reliability and meets the minimum accepted threshold of 0.7 since most variances are due to signal and not noise. We feel comfortable that the measure is capturing data in a reliable fashion to measure the performance of physicians.

_____

2b2. VALIDITY TESTING

2b2.1. What level of validity testing was conducted? (*may be one or both levels*)

☐ **Critical data elements** (*data element validity must address ALL critical data elements*)

☒ **Performance measure score**

☒ **Empirical validity testing**

☒ **Systematic assessment of face validity of performance measure score as an indicator** of quality or resource use (*i.e., is an accurate reflection of performance on quality or resource use and can distinguish good from poor performance*)

2b2.2. For each level of testing checked above, describe the method of validity testing and what it tests (*describe the steps—do not just name a method; what was tested, e.g., accuracy of data elements compared to authoritative source, relationship to another measure as expected; what statistical analysis was used)*

Empirical Validity:

Method of testing construct validity: At the health plan level, we tested for construct validity by exploring whether Comprehensive Diabetes Care: BP Control <140/90 mm Hg (# 0061) correlated with the *Controlling High Blood Pressure* measure #0018). The Controlling High Blood Pressure measure assesses the percentage of adults 18-85 years with a confirmed diagnosis of hypertension whose blood pressure was adequately controlled (<140/90 mm Hg). Correlations were also tested with the following NQF endorsed diabetes measures:

- *HbA1c Testing (#0057)*: The percentage of adults 18-75 with diabetes that had an HbA1c test performed during the measurement year
- *HbA1c Poor Control (> 9.0%) (#0059)*: The percentage of adults 18-75 with diabetes whose most recent HbA1c level is >9%
- *HbA1c Control (<8.0%):(#0575)* The percentage of adults 18-75 with diabetes whose most recent HbA1c level is <8%
- *Eye Exam(#0055):* The percentage of adults 18-75 with diabetes that had an eye screening for diabetic retinal disease during the measurement year
- *Medical Attention for Nephropathy(#0062)*: The percentage of adults 18-75 with diabetes that had a nephropathy screening test or evidence of nephropathy during the measurement year

We specifically hypothesized that Blood Pressure Control (<140/90 mm Hg) will be positively correlated with the Controlling High Blood Pressure measure (i.e. plans that have high rates of blood pressure

control in the diabetes population will have high rates of blood pressure control in the general population).

The Diabetes Recognition Program uses BP >=140/90 mm Hg to measure poor blood pressure control. *Lower score = Better quality. We tested correlations between BP >=140/90 mm Hg with other measures in the Diabetes Recognition Program. These measures included HbA1c Poor Control (>9%), HbA1c Control (<8%), Eye Exam, Medical Attention for Nephropathy, Foot Exam and BP <130/80 mm Hg. The DRP uses BP <130/80 mm Hg as a measure for good blood pressure control.

To test these correlations we used a Person correlation test. This test estimates the strength of the linear association between two continuous variables; the magnitude of correlation ranges from -1 and +1. A value of 1 indicates a perfect linear dependence in which increasing values on one variable is associated with increasing values of the second variable. A value of 0 indicates no linear association. A value of -1 indicates a perfect linear relationship in which increasing values of the first variable is associated with decreasing values of the second variable.

Systematic Assessment of Face Validity:

Health Plan Level

Method of Assessing Face Validity: NCQA has identified and refined measure management into a standardized process called the HEDIS measure life cycle.

STEP 1: NCQA staff identifies areas of interest or gaps in care. Clinical expert panels (MAPs—whose members are authorities on clinical priorities for measurement) participate in this process. Once topics are identified, a literature review is conducted to find supporting documentation on their importance, scientific soundness and feasibility. This information is gathered into a work-up format. Refer to What Makes a Measure "Desirable"? The work-up is vetted by NCQA's Measurement Advisory Panels (MAPs), the Technical Measurement Advisory Panel (TMAP) and the Committee on Performance Measurement (CPM) as well as other panels as necessary.

STEP 2: Development ensures that measures are fully defined and tested before the organization collects them. MAPs participate in this process by helping identify the best measures for assessing health care performance in clinical areas identified in the topic selection phase. Development includes the following tasks: (1) Prepare a detailed conceptual and operational work-up that includes a testing proposal and (2) Collaborate with health plans to conduct field-tests that assess the feasibility and validity of potential measures. The CPM uses testing results and proposed final specifications to determine if the measure will move forward to Public Comment.

STEP 3: Public Comment is a 30-day period of review that allows interested parties to offer feedback to NCQA and the CPM about new measures or about changes to existing measures.
NCQA MAPs and technical panels consider all comments and advise NCQA staff on appropriate recommendations brought to the CPM. The CPM reviews all comments before making a final decision about Public Comment measures. New measures and changes to existing measures approved by the CPM will be included in the next HEDIS year and reported as first-year measures.

STEP 4: First-year data collection requires organizations to collect, be audited on and report these measures, but results are not publicly reported in the first year and are not included in NCQA's State of Health Care Quality, Quality Compass or in accreditation scoring. The first-year distinction guarantees

that a measure can be effectively collected, reported and audited before it is used for public accountability or accreditation. This is not testing—the measure was already tested as part of its development—rather, it ensures that there are no unforeseen problems when the measure is implemented in the real world. NCQA's experience is that the first year of large-scale data collection often reveals unanticipated issues. After collection, reporting and auditing on a one-year introductory basis, NCQA conducts a detailed evaluation of first-year data. The CPM uses evaluation results to decide whether the measure should become publicly reportable or whether it needs further modifications.

STEP 5: Public reporting is based on the first-year measure evaluation results. If the measure is approved, it will be publically reported and may be used for scoring in accreditation.

Step 6: Evaluation is the ongoing review of a measure's performance and recommendations for its modification or retirement. Every measure is reviewed for reevaluation at least every three years. NCQA staff continually monitors the performance of publicly reported measures. Statistical analysis, audit result review and user comments through NCQA's Policy Clarification Support portal contribute to measure refinement during re-evaluation. Information derived from analyzing the performance of existing measures is used to improve development of the next generation of measures.

Each year, NCQA prioritizes measures for re-evaluation and selected measures are researched for changes in clinical guidelines or in the health care delivery systems, and the results from previous years are analyzed. Measure work-ups are updated with new information gathered from the literature review, and the appropriate MAPs review the work-ups and the previous year's data. If necessary, the measure specification may be updated or the measure may be recommended for retirement. The CPM reviews recommendations from the evaluation process and approves or rejects the recommendation. If approved, the change is included in the next year's HEDIS Volume 2.

Expert Participation

This measure was tested for face validity with input from three expert panels. Updated guidelines from the eighth Joint National Committee (JNC 8) in December 2013 motivated the measure developers to re-evaluate blood pressure control indicators in the Comprehensive Diabetes Care measure set, which previously included two blood pressure control measures (<140/80 mm Hg-not NQF endorsed) and (<140/90 mm Hg-NQF #0061). The measure developers solicited input from stakeholders to ascertain whether a blood pressure level <140/90 mm Hg is the appropriate blood pressure control level for patients with diabetes. The blood pressure control <140/80 mm Hg measure (not NQF endorsed) was retired following stakeholder input. Our stakeholders supported maintaining the blood pressure control <140/90 mm Hg measure (NQF #0061) as the only blood pressure threshold for quality care in patients with diabetes to align with the JNC 8 guidelines. See evidence form 1a.4.2 for JNC 8 guideline recommendations.

We list an overview of each panel here. Please refer to Ad.1 in the submission form for the names and affiliation of experts in each panel.

4. Diabetes Measurement Advisory Panel includes 17 clinicians with expertise in endocrinology, podiatry, nephropathy, ophthalmology, cardiology, internal medicine/family medicine, pediatric endocrinology pharmacy and diabetes education. Other stakeholders include one consumer advocate and one health advisory consultant with expertise in diabetes treatment.
5. Cardiovascular Measurement Advisory Panel includes eight physicians and one nurse with expertise in cardiovascular health and quality measurement.

6. NCQA's Committee on Performance Measurement (CPM) oversees the evolution of the measurement set and includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 16 members. The CPM is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of performance measures. CPM members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and the science of measurement.

Diabetes Recognition Program

Measures are tested for face validity with three panels of experts. The Diabetes Recognition Program (DRP) Advisory Committee included 7 experts in diabetes care including representation by clinicians, health plans, integrated health systems and research organizations; the Diabetes Measurement Advisory Panel included 16 experts in diabetes care, including representation by consumers, health plans, health care providers and policy makers. NCQA's Clinical Programs Committee (CPC) oversees the evolution of NCQA's recognition programs and related measures including the Diabetes Recognition Program, the Heart/Stroke Recognition Program, the Patient Centered Medical Home and Patient-Centered Specialty Practice Recognition Program, among others. The CPC includes representation by purchasers, consumers, health plans, health care providers and policy makers. This panel is made up of 17 members. The CPC is organized and managed by NCQA and reports to the NCQA Board of Directors and is responsible for advising NCQA staff on the development and maintenance of clinical recognition programs. CPC members reflect the diversity of constituencies that performance measurement serves; some bring other perspectives and additional expertise in quality management and in improvement science.

ICD-10 CONVERSION:

Goal was to convert this measure to a new code set, fully consistent with the intent of the original measure.

Steps in ICD-9 to ICD-10 Conversion Process

1. NCQA staff identify ICD-10 codes to be considered based on ICD-9 codes currently in measure. Use GEM to identify ICD-10 codes that map to ICD-9 codes. Review GEM mapping in both directions (ICD-9 to ICD-10 and ICD-10 to ICD-9) to identify potential trending issues.
2. NCQA staff identify additional codes (not identified by GEM mapping step) that should be considered. Using ICD-10 tabular list and ICD-10 Index, search by diagnosis or procedure name for appropriate codes.
3. NCQA HEDIS Expert Coding Panel review NCQA staff recommendations and provide feedback.
4. As needed, NCQA Measurement Advisory Panels perform clinical review. Due to increased specificity in ICD-10, new codes and definitions require review to confirm the diagnosis or procedure is intended to be included in the scope of the measure. Not all ICD-10 recommendations are reviewed by NCQA MAP; MAP review items are identified during staff conversion or by HEDIS Expert Coding Panel.
5. Post ICD-10 code recommendations for public review and comment.
6. Reconcile public comments. Obtain additional feedback from HEDIS Expert Coding Panel and MAPs as needed.
7. NCQA staff finalize ICD-10 code recommendations.

Tools Used to Identify/Map to ICD-10

All tools used for mapping/code identification from CMS ICD-10 website (http://www.cms.gov/Medicare/Coding/ICD10/2012-ICD-10-CM-and-GEMs.html).
GEM, ICD-10 Guidelines, ICD-10-CM Tabular List of Diseases and Injuries, ICD-10-PCS Tabular List.

2b2.3. What were the statistical results from validity testing? (*e.g., correlation; t-test*)

Empirical Validity

Health Plan Level

Correlations among Comprehensive Diabetes Care measure indicators and Controlling High Blood Pressure measure 2014

Quality Measure	Pearson Correlation Coefficients						
	Controlling High Blood Pressure (0018)	CDC HbA1c Testing	CDC HbA1c Poor Control (>9%)	CDC HbA1c Control (<8%)	CDC Eye Exam	CDC Med Att Diabetic Neph.	CDC Blood Pressure Ctrl <140/90
CDC Blood Press Ctrl <140/90 mm Hg (Commercial)	**0.80878**	0.62849	**-0.85003**	**0.80244**	0.62712	0.61774	1
CDC Blood Press Ctrl <140/90 mm Hg (Medicaid)	**0.82736**	0.5316	**-0.76551**	**0.75364**	0.48903	0.34173	1
CDC Blood Press Ctrl <140/90 mm Hg (Medicare)	**0.76084**	0.4808	-0.64626	0.62281	0.53324	0.31794	1

Note: All correlations are significant at p<.05

Diabetes Recognition Program (Clinician and Practice Combined)

Correlations among Diabetes Recognition Program measures-2013

Quality Measure	Pearson Correlation Coefficients						
	BP >=140/90 mm Hg	HbA1c >9%	HbA1c <8%	BP <130/80 mm Hg	Eye Exam	Nephropathy	Foot Exam
BP >=140/90 mm Hg	1	-0.05974	0.07575	**-0.57785**	-0.21321	-0.11448	-0.05757

Systematic Assessment of Face Validity

Expert Panels

Our expert panels unanimously supported a blood pressure control threshold of <140/90 mm Hg to measure quality care in patients with diabetes.

2b2.4. What is your interpretation of the results in terms of demonstrating validity? (i*.e., what do the results mean and what are the norms for the test conducted?*)

Empirical Validity

Health Plan

Correlations testing suggests that the CDC Blood Pressure Control **<**140/90 mm Hg has a strong positive correlation with the Controlling High Blood Pressure measure, as hypothesized. Health plans that perform well on the CDC Blood Pressure Control **<**140/90 mm Hg measure also have a strong positive correlation with the HbA1c Control <8% measure and a strong negative correlation with the HbA1c Poor Control **>**9% measure, which was expected.

Diabetes Recognition Program

Correlations testing suggests that there is a moderate negative correlation between the BP **>=** 140/90 mm Hg and BP **<**130/80 mm Hg. The Diabetes Recognition Program includes BP **>=**140/90 mm Hg as a poor control measure and BP **<**130/80 mm Hg as a good control measure.

Systematic Assessment of Face Validity

Expert Panels

Experts agreed that the JNC 8 recommendations are the authoritative source in setting blood pressure control goals for patients with diabetes.

2b3. EXCLUSIONS ANALYSIS

NA ☒ no exclusions — *skip to section 2b4*

2b3.1. Describe the method of testing exclusions and what it tests (*describe the steps—do not just name a method; what was tested, e.g., whether exclusions affect overall performance scores; what statistical analysis was used*)

2b3.2. What were the statistical results from testing exclusions? (*include overall number and percentage of individuals excluded, frequency distribution of exclusions across measured entities, and impact on performance measure scores*)

2b3.3. What is your interpretation of the results in terms of demonstrating that exclusions are needed to prevent unfair distortion of performance results? (*i.e., the value outweighs the burden of increased data collection and analysis. Note: **If patient preference is an exclusion**, the measure must be specified so that the effect on the performance score is transparent, e.g., scores with and without exclusion*)

2b4. RISK ADJUSTMENT/STRATIFICATION FOR OUTCOME OR RESOURCE USE MEASURES

If not an intermediate or health outcome, or PRO-PM, or resource use measure, skip to section 2b5.

2b4.1. What method of controlling for differences in case mix is used?

☒ **No risk adjustment or stratification**

☐ **Statistical risk model with** Click here to enter number of factors **risk factors**

☐ **Stratification by** Click here to enter number of categories **risk categories**
☐ **Other,** Click here to enter description

2b4.2. If an outcome or resource use measure is not risk adjusted or stratified, provide rationale and analyses to demonstrate that controlling for differences in patient characteristics (case mix) is not needed to achieve fair comparisons across measured entities.

2b4.3. Describe the conceptual/clinical and statistical methods and criteria used to select patient factors used in the statistical risk model or for stratification by risk (*e.g., potential factors identified in the literature and/or expert panel; regression analysis; statistical significance of p<0.10; correlation of x or higher; patient factors should be present at the start of care and not related to disparities*)

2b4.4. What were the statistical results of the analyses used to select risk factors?

2b4.5. Describe the method of testing/analysis used to develop and validate the adequacy of the statistical model or stratification approach (*describe the steps—do not just name a method; what statistical analysis was used*)

Provide the statistical results from testing the approach to controlling for differences in patient characteristics (case mix) below.
If stratified, skip to 2b4.9

2b4.6. Statistical Risk Model Discrimination Statistics (*e.g., c-statistic, R-squared*)**:**

2b4.7. Statistical Risk Model Calibration Statistics (*e.g., Hosmer-Lemeshow statistic*):

2b4.8. Statistical Risk Model Calibration – Risk decile plots or calibration curves:

2b4.9. Results of Risk Stratification Analysis:

2b4.10. What is your interpretation of the results in terms of demonstrating adequacy of controlling for differences in patient characteristics (case mix)? (i.*e., what do the results mean and what are the norms for the test conducted*)

2b4.11. Optional Additional Testing for Risk Adjustment (*not required, but would provide additional support of adequacy of risk model, e.g., testing of risk model in another data set; sensitivity analysis for missing data; other methods that were assessed*)

2b5. IDENTIFICATION OF STATISTICALLY SIGNIFICANT & MEANINGFUL DIFFERENCES IN PERFORMANCE

2b5.1. Describe the method for determining if statistically significant and clinically/practically meaningful differences in performance measure scores among the measured entities can be identified (*describe the steps—do not just name a method; what statistical analysis was used? Do not just repeat the information provided related to performance gap in 1b)*

To demonstrate meaningful differences in performance, NCQA calculates an inter-quartile range (IQR) for each indicator. The IQR provides a measure of the dispersion of performance. The IQR can be interpreted as the difference between the 25th and 75th percentile on a measure. To determine if this difference is statistically significant, NCQA calculates an independent sample t-test of the performance difference between two randomly selected plans at the 25th and 75th percentile. The t-test method calculates a testing statistic based on the sample size, performance rate, and standardized error of each plan. The test statistic is then compared against a normal distribution. If the p value of the test statistic is less than .05, then the two plans' performance is significantly different from each other. Using this method, we compared the performance rates of two randomly selected plans, one plan in the 25th percentile and another plan in the 75th percentile of performance. We used these two plans as examples of measured entities. However, the method can be used for comparison of any two measured entities

2b5.2. What were the statistical results from testing the ability to identify statistically significant and/or clinically/practically meaningful differences in performance measure scores across measured entities? (e.g., *number and percentage of entities with scores that were statistically significantly different from mean or some benchmark, different from expected; how was meaningful difference defined*)

Health Plan Level: HEDIS 2014 Variation in Performance across Health Plans

	N	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
Commercial	342	62%	13%	49%	56%	63%	71%	75%	14%	<0.05
Medicaid	209	60%	14%	46%	53%	61%	70%	75%	16%	<0.05
Medicare	489	65%	11%	54%	59%	65%	72%	77%	13%	<0.05

N= Number of plans reporting
IQR: Interquartile range
p-value: P-value of independent samples t-test comparing plans at the 25th percentile to plans at the 75th percentile.

Physician Level: Diabetes Recognition Program 2013 Variation in Performance across Physicians (Clinician and Practice Combined)

*Please Note: The Diabetes Recognition Program measure is BP **>=** 140/90 mm Hg. **Lower score = Better quality**

	N	Avg.	SD	10th	25th	50th	75th	90th	IQR	p-value
DRP	2,477	19%	10%	7%	12%	18%	26%	32%	14%	<0.05

N= total number of clinicians reporting data for recognition including clinicians in group practices

2b5.3. What is your interpretation of the results in terms of demonstrating the ability to identify statistically significant and/or clinically/practically meaningful differences in performance across measured entities? (i*.e., what do the results mean in terms of statistical and meaningful differences?*)

Health Plan Level
The results indicate there is a 13-16% gap in performance between the 25th and 75th performing plans. The largest performance gap (16%) is in Medicaid plans. The difference between the 25th and 75th percentile is statistically significant for all product lines (Commercial, Medicaid, Medicare). There is also a 23-29% gap in performance between the 10th and the 90th performing plans. Overall, results suggest there are meaningful differences in performance and there is an opportunity for improvement.

Diabetes Recognition Program
Results indicate there is a 14% gap in performance between the 25th (better performance for this measure in the DRP program) and 75th performing clinicians in the Diabetes Recognition Program. This suggests there were 347 more clinicians in the 25th percentile than in the 75th percentile. The difference between the 25th and 75th percentile is statistically significant for clinicians submitting data on this measure to the Diabetes Recognition Program.

2b6. COMPARABILITY OF PERFORMANCE SCORES WHEN MORE THAN ONE SET OF SPECIFICATIONS
***If only one set of specifications, this section can be skipped*.**

Note: *This criterion is directed to measures with more than one set of specifications/instructions (e.g., one set of specifications for how to identify and compute the measure from medical record abstraction and a different set of specifications for claims or eMeasures). It does not apply to measures that use more than one source of data in one set of specifications/instructions (e.g., claims data to identify the denominator and medical record abstraction for the numerator).* ***If comparability is not demonstrated, the different specifications should be submitted as separate measures.***

2b6.1. Describe the method of testing conducted to demonstrate comparability of performance scores for the same entities across the different data sources/specifications (*describe the steps—do not just name a method; what statistical analysis was used*)
N/A

2b6.2. What were the statistical results from testing comparability of performance scores for the same entities when using different data sources/specifications? (*e.g., correlation, rank order*)
N/A

2b6.3. What is your interpretation of the results in terms of demonstrating comparability of performance measure scores for the same entities across the different data sources/specifications? (i.*e., what do the results mean and what are the norms for the test conducted*)
N/A

2b7. MISSING DATA ANALYSIS AND MINIMIZING BIAS

2b7.1. Describe the method of testing conducted to identify the extent and distribution of missing data (or nonresponse) and demonstrate that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias (*describe the steps—do not just name a method; what statistical analysis was used*)

This measure is collected with a complete sample, there is no missing data on this measure.

2b7.2. What is the overall frequency of missing data, the distribution of missing data across providers, and the results from testing related to missing data? (*e.g., results of sensitivity analysis of the effect of various rules for missing data/nonresponse; if no empirical sensitivity analysis, identify the approaches for handling missing data that were considered and pros and cons of each*)
This measure is collected with a complete sample, there is no missing data on this measure.

2b7.3. What is your interpretation of the results in terms of demonstrating that performance results are not biased due to systematic missing data (or differences between responders and nonresponders) and how the specified handling of missing data minimizes bias**?** (i.*e., what do the results mean in terms of supporting the selected approach for missing data and what are the norms for the test conducted; if no empirical analysis, provide rationale for the selected approach for missing data*)
This measure is collected with a complete sample, there is no missing data on this measure.